EAST ASIAN ANCESTRY IN INDIA

Gyaneshwer Chaubey

ABSTRACT
It has been suggested that Indian populations constitute the second largest diversity of human populations just after Africa. Based on the mitochondrial DNA evidence it was suggested that the prehistoric South Asia (including Southeast Asia) carried half of the world populations 20-40 thousands years ago. Such scenario led some to conclude a complex and deep demographic history of the subcontinent with minor gene flow from East and the West. The relatedness of India to Central Asian, European, Middle Eastern, the Caucasus and to the East/Southeast Asians has been suggested, but none of the study has estimated the East/Southeast Asian ancestry among the extent population of India. Here the analysis of genome wide data on Indian and East/Southeast Asian demonstrated their restricted distinctive ancestry in India mainly running along the foothills of Himalaya and northeastern part. Moreover, we have also identified the consistency of East/Southeast Asian ancestry over the population history of the subcontinent leading the entry of Austroasiatic and Tibeto-Burman languages.

INTRODUCTION
Studies on Indian populations by various disciplines yielded contrasting results (Chaubey et al., 2007; Petraglia et al., 2007; Anthony 2009; Reich et al., 2009; Xing et al., 2010; Metspalu et al., 2011; Moorjani et al., 2013). In light of limited archaeological evidence and the limitations of linguistics, molecular anthropology is another independent line of evidence which may resolve such complexity; however, different genetic systems are also not in congruent. The mitochondrial DNA (mtDNA) overwhelmingly support an early arrival of modern humans to the subcontinent with later minor admixture from East and West (Metspalu et al., 2004; Thangaraj et al., 2006; Metspalu et al., 2004; Sun et al., 2006; Chandrasekar et al., 2009). The Y chromosome studies have identified frequent autochthonous lineages (haplogroups C5, F*, H, L and R2) and also lineages with either Middle Eastern (haplogroup J2) or East/Southeast Asian (haplogroup O2a and O3a) or of unknown origins (haplogroup R1a) (Sahoo et al., 2006; Sengupta et al., 2006; Underhill et al., 2015; Singh et al., 2016; Chaubey et al., 2016). On the other hand, the autosomal

---

Dr. Gyaneshwer Chaubey, Evolutionary Biology Group, Estonian Biocentre, Riia23, Tartu, Estonia-51010, E-mail: gyanc@ebc.ee
studies have identified two major components: one largely restricted to the subcontinent, whilst second is shared with Central Asia, the Caucasus, Middle East and Europe (Reich et al., 2009; Metspalu et al., 2011). This striking sharing was in coherence with the historic linguistics (Anthony, 2009), nevertheless the haplotype based analysis ruled it out by showing its dispersal timeline to be before Neolithic (Metspalu et al., 2011).

Along with these two major components, studies on autosomes also showed minor components restricted to some populations which are however prevalent in East and West Eurasia (Chaubey et al., 2011; Chaubey et al., 2015). One of these components is present among Indian Austroasiatic (Munda) speakers which are thought to be associated with the arrival of Austroasiatic speakers from Southeast Asia to India (Chaubey et al., 2011). The arrival of Austroasiatic speakers from Southeast Asia to India is also attested with the frequent Southeast Asian haplogroup O2a (M95) among Indian Munda speakers (Chaubey et al., 2011). The virtual absence of East/Southeast Asian specific maternal lineages among Munda speakers helped researchers to conclude that the Austroasiatic (Munda) migration to India was mainly male mediated (Chaubey et al., 2011). Another Austroasiatic (Khasi-Khamic) speaker of India, Khasi doesn’t show sex specific admixture as Munda (Reddy et al., 2007). Tibeto-Burman speakers of India also corroborate their affinity with the East/Southeast Asia (Metspalu et al., 2004). The presence of distinct East/Southeast Asian specific ancestry among Indian Austroasiatic and Indian Tibeto-Burmans indicate their different population histories in the subcontinent (Chaubey et al., 2014).

It is clear that most of the East/Southeast Asian ancestry to India is arrived via language dispersal (Diamond and Bellwood, 2003) and other local demographic histories (Thangaraj et al., 2008; Chaubey et al., 2014), but it is a question that at what extant it is diffused into the subcontinent beyond its linguistic boundaries? To investigate this issue we have calculated the East/Southeast Asian ancestry among thirty Indian populations belong to different linguistic groups and two Negrito populations from Andaman Island.

MATERIAL AND METHODS

In this study, we have merged samples from three different studies (HUGO Pan-Asian SNP Consortium et al., 2009; Reich et al., 2009; International HapMap 3 Consortium et al., 2010). Average IBS (identity by state) was calculated (Purcell et al., 2007), and related individuals up to three generations were removed from the analysis. We used PLINK 1.07 (Purcell et al., 2007) to filter the combined data set to include only SNPs on the 22 autosomal chromosomes with minor allele frequency >1% and genotyping success >99%. Our combined dataset had data for 12,622 SNPs, after excluding SNPs unique to any of the three platforms and SNPs from mtDNA and X and Y chromosomes. To calculate the East Asian specific ancestry among Indian populations we have used the $f_4$ ratio estimation test (Patterson et al., 2012). We used European-CEU as an out group and calculated the Han ancestry among
the Indian populations; East Asian ancestry = $f_4^{(Japan, Kurumba; X, CEU)}/f_4^{(Japan, Kurumba; Han, CEU)}$, where $X$ represent Indian populations. We plotted the % of ancestry over the geographical map of India (Fig. 1).

**RESULTS AND DISCUSSION**

India is linguistically, genetically, culturally and geographically a highly heterogeneous country. The highly ripped caste and tribal groups of India are unique among all known societies in human history (Chaubey et al., 2007). This uniqueness is the result of long term high effective population size and admixture to and from East and West. In the present study we have quantified the East/Southeast Asian gene flow among several populations of India belonging to different language groups.

We observed a high frequency of East/Southeast Asian specific ancestry over the lower foothills of Himalaya and towards northeast India (Fig. 1). Nishi, a Tibeto-Burman population from Arunanchal Pradesh carried the highest amount (93%).
of East/Southeast Asian ancestry, followed by the Ao-Naga (92%) from Nagaland belongs to the same language family. The highest frequency of East/Southeast Asian ancestry among Tibeto-Burman populations support their recent arrival from East (Cordaux et al., 2004; Metspalu et al., 2004). Notably, the Negrito populations from Andaman Island also showed a high level of East/Southeast Asian ancestry (Fig. 1). Our previous analyses on these two populations have observed three distinct ancestry components and have suggested a deep common ancestry of Andaman Negrito with the Melanesia, Malaysian Negrito and South Asia (Chaubey and Endicott, 2013). The Han ancestry measured in Andaman Negrito is probably partially capturing both the Melanesian and Malaysian Negrito ancestry.

The Transitional and Munda groups were also harbouring a substantial amount of East/Southeast Asian ancestry (Fig. 1). Consistent with the higher genome sharing with East/Southeast Asia, South Munda (Kharia) constituted 21% of the East/Southeast Asian ancestry, in comparison with North Munda (Santhal 13%). Saharia a Transitional group, who presently speak Indo-European language, carried 15% of the East/Southeast Asian ancestry. Satnami another transitional group speaking Indo-European language carried 5% of the East/Southeast Asian ancestry (Fig. 1). It is interesting to note that the Transitional populations share the same geography as their Austroasiatic neighbours and it is highly likely that the East/Southeast Asian ancestry among them is a result of gene flow or language shift.

Among the large number of Indo-European and Dravidian speakers the East/Southeast Asian component was largely absent except in Tharu and Chenchu (Fig. 1). The case of Tharu ancestry have been already discussed elsewhere in detail (Thangaraj et al., 2008; Chaubey et al., 2014), whereas there is no historic information about language shift among Chenchu, who presently speak Dravidian language. The presence of 4% (though insignificant Z < 2) of East/Southeast Asian ancestry among Chenchu is intriguing. It is likely that they may have received minor gene flow from the Austroasiatic populations living in neighbouring state, which is evident in their maternal profile (Endicott et al., 2006). Moreover, the East/Southeast Asian specific Y chromosome lineages were not observed among Chenchu in haploid DNA analysis (Kivisild et al., 2003).

In conclusions, considering the overall distribution of East/Southeast Asian ancestry in India, it is clear that due to the high level of endogamy the diffusion of this ancestry component is limited with the exception of the populations who have known introgression from the population associated with the Austroasiatic or Tibeto-Burman e.g. Tharu, Saharia and Satnami. The spatial distribution of this component largely mimics the spread of Austroasiatic, Tibeto-Burman languages; Y chromosomal haplogroups O2a and O3a; and autosomal EDAR 370A variant (Chaubey et al., 2011). Therefore, taking into consideration the demographic histories of the Indian populations, the distribution of East/Southeast Asian ancestry in India predominantly follow linguistic landscape model followed by ‘isolation by distance’ model.
ACKNOWLEDGEMENTS

This study is supported by Estonian Personal grants PUT-766 (GC). GC also acknowledges the financial support from European Union European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre and University of Tartu by Tartu University grant (PBGMR06901), and Estonian Institutional Research grants IUT24-1.

REFERENCES

Anthony DW 2009. The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton University Press.


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007; 81(3): 559-575.


